The Fermilab Computing Farms in 2001-2002

Merina Albert, Lisa Giacchetti, Margaret Greaney, Terry Jones, Tanya Levshina, Jeff Mack, Igor Mandrichenko, Stan Naymola, Ray Pasetes, Ken Schumacher, Karen Shepelak, Joseph Syu, Steven Timm, Stephen Wolbers

Version 2.2
April 24, 2003

Introduction

The Fermilab computing farms grew substantially in 2001 and 2002. This reflected primarily the CDF and D0 computing demand increase as run 2 began and the two detectors and the accelerator performed steadily better, leading to more data and a greater demand for reconstruction computing. In addition, the "fixed-target" farms evolved away from the old model of direct tape input and output to a system that uses Enstore[1] and dcache[2] (network-based) as the I/O mechanism. This was part of a more general trend away from a large server used for all of the I/O and many common services and many workers to a model with many smaller systems serving as I/O systems with distributed disk storage. Other major technological achievements include the use of dfarm[3], a disk caching mechanism, throughout the farms, the upgrades of FBSNG[4], NGOP[5] for monitoring the farms systems, and generally a more sophisticated management of the machines that constitute these farms. The growth in systems was quite substantial, from a total of 314 dual PCs in early 2001 to 633 duals in early 2003. This does not include the farms that were purchased for CDF and D0 analysis – the CAF and the CAB, nor does it include CMS PC systems. Those systems are sufficiently different in use (and are used differently by the collaborations) that they are treated separately. In future versions of this memo it is likely that all of the systems will be covered.

2001 and 2002 in review

The "fixed-target" or "general" PC farm has seen a constant change of hardware and customer base over the two years 2001 and 2002. The systems currently consist of

50 dual-Pentium III 500 MHz systems, 40 dual-Pentium III 1 GHz systems, and 16 dual-Pentium III 1.26 GHz systems. The 16-1.26 GHz systems were purchased with Japanese funds to help KTeV get access to the amount of CPU required for its analysis. All users have access to all nodes with KTeV having some priority on some of the systems. During the past two years the farm has been reorganized from two logical and distinct farms into one farm with two I/O nodes (fnsfh and fnsfo). E871 and KTeV are the only remaining FT99 experiments using the farms. Current users include BTeV, miniBooNE, MINOS, NUMI, Auger, SDSS, accelerator simulations, theory calculations and simulations, and NLC detector simulations. This reflects the changing user community as well as the push to include others at the lab that can make use of this facility for calculations that cannot be easily accomplished at the laboratory. The size of the facility is based partially on demand and it is not so hard to increase it by modest to medium amounts if the demand warrants it.

The CDF farm has grown and has been modified many times over the period 2001-2002 to reflect changes in demand, in I/O system architecture changes, use of local disk as staging areas (dfarm), among other things. Currently the system has 151 nodes, 23 dual-Pentium III 800 MHz systems, 64 dual-Pentium III 1 GHz systems, 32 dual-Pentium III 1.26 GHz systems and 32 dual Athlon 2000+ (1.67 GHz) systems. 16 of the PCs are Gbit connected to the CDF farms switch and allows them to serve as fast I/O systems to the CDF Enstore tapedrives. 3 servers are also part of the farm – 1 SGI O2200, and 2 PCs. These systems handle NIS and NFS, mySQL database serving, FBSNG and related software and web serving. (The 50 500 MHz nodes were decommissioned in January, 2003.)

The D0 farm has grown very rapidly over the past two years, reflecting the strong demand for CPU time. Currently the D0 farm consists of 342 nodes, 20 dual-Pentium III 500 MHz PCs, 50 dual-Pentium III 800 MHz PCs, 32 dual-Pentium III 1 GHz PCs, and 240 dual-Athlon 2000+ (1.67 GHz) PCs. I/O and various server functions are handled by an SGI O2000. A subset of the PCs are used for I/O as well.

The CMS farm consists of 136 systems, 40 dual-Pentium III 750 MHz PCs, 32 dual-Pentium III 1 GHz PCs, and 64 dual-Athlon 2000+ (1.67 GHz) PCs. The CMS farms have grown steadily to meet the requirements of the various data challenges.

CPU Utilization

Tables 1 and 2 provide a summary of CPU time (in SpecInt95 units) for the CDF, D0 and general farms. A plot of the total CPU utilization, going back to 1991, is shown in figure 1. First, it is interesting to see the increase in computing power delivered during the past 12 years. In units of SpecInt95 the increase is from approximately 10 to about 35,000 – an increase of a factor of 3500! When one factors in that the number of CPUs increased from 25 in early 1991 to 1266 in 2003 this is not quite so impressive an increase. But given that budgets have been reasonably stable over the period this still represents a large increase in CPU consistent with Moore's law translated to price performance.
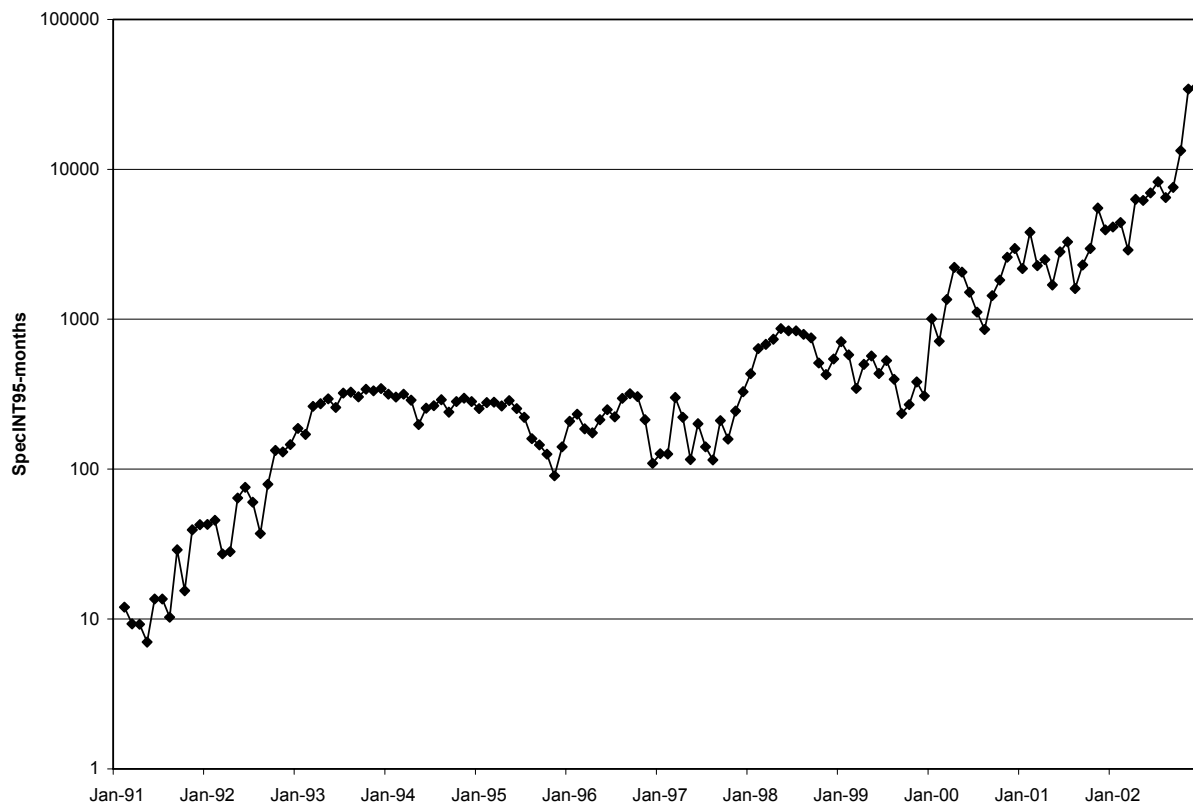
Figure 1. Farms Usage History, 1991-2002

Table 1. Total CPU use on the Farms: 2001-2002

	SpecINT95-months
January 2001	2,177
February 2001	3,804
March 2001	2,281
April 2001	2,495
May 2001	1,695
June 2001	2,813
July 2001	3,285
August 2001	1,603
September 2001	2,300
October 2001	2,965
November 2001	5,523
December 2001	3,944
January 2002	4,124
February 2002	4,426
March 2002	2,897
April 2002	6,301
May 2002	6,189
June 2002	6,963
July 2002	8,269
August 2002	6,476
September 2002	7,575
October 2002	13,309
November 2002	27,099
December 2002	35,121

Table 2. Farms, 2001-2002
SpecINT95-months

Experiment	2001	2002
Auger	0	5390
BTeV	2495	714
CDF	6030	20811
D0	6678	80679
E781	2319	0
E835	17	12
E871	3840	6220
E898	338	772
KTeV	3920	6064
LCD	0	113
NUMI/MINOS	3068	1933
Theory	58	25
SDSS	1161	1725
Total	29924	124458

The next set of plots (Figure 2,3,4) show the usage during the past 2 years by experiment and by month. Note that there is a significant increase starting in late 2002. This was due to the increase in the farms capacity generated by the addition of 240 nodes (for D0) and smaller increases for CDF and the general farms. In addition, the demand for processing data increased as Run 2 matured. CDF embarked on a complete reprocessing of all 2002 data in October, 2002. Both CDF and D0 wished to process data for the winter conferences.

Figure 2. Farms Usage by Experiment, 2001-2002.
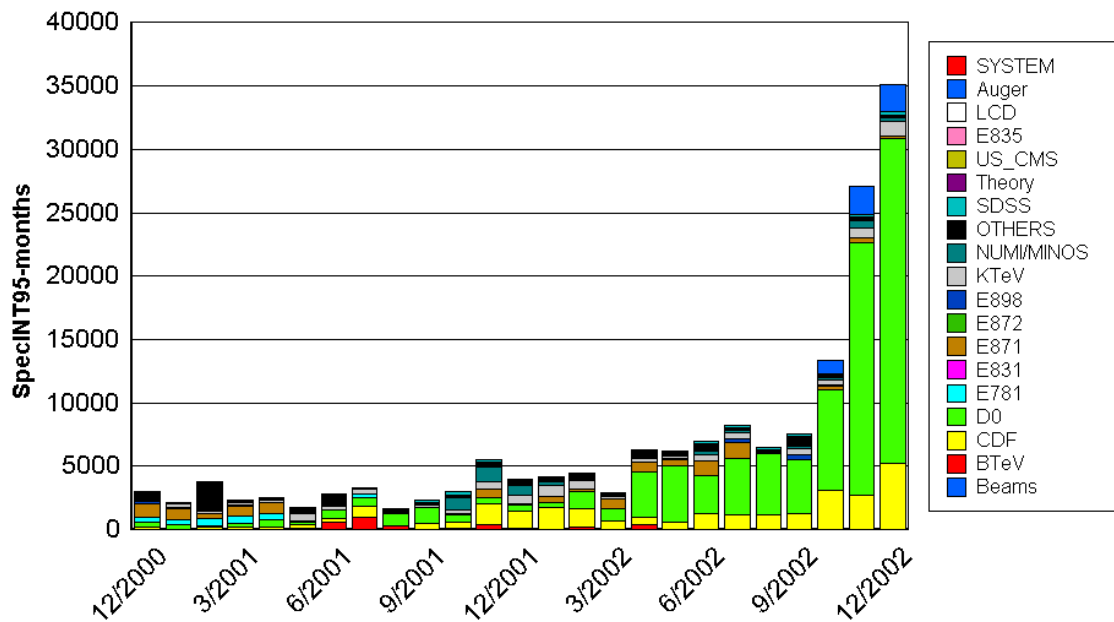
Other experiments made good use of the farms during this period. Many of the experiments that ran in the 1996-97 and 1999 fixed-target runs processed data or embarked on Monte Carlo and final analysis on the farms during this period. This

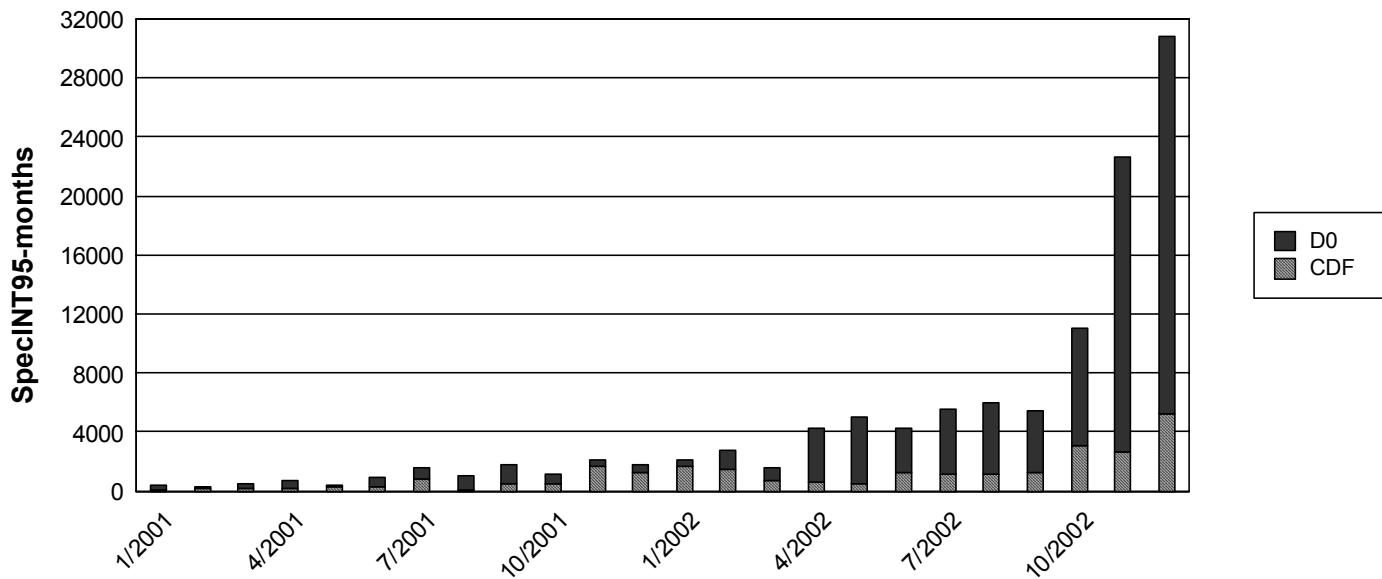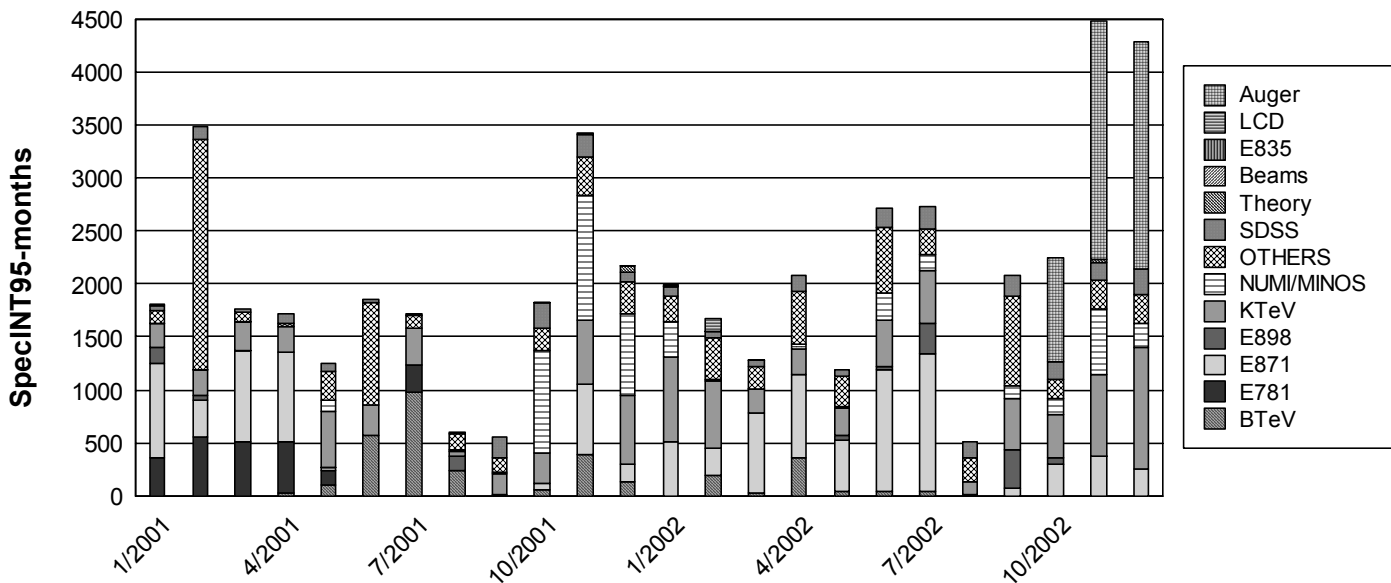Figure 3. Farms Usage Total for CDF and D0.

Figure 4. Farms Usage Total for All Other Experiments and Groups.

includes E781, E835, E871, E872 and KTeV. The neutrino experiments miniBooNE (E898) and MINOS (also listed and combined with NUMI) ramped up use of the farms for simulation and data processing. SDSS uses the farms for processing and reprocessing of data. BTeV, Auger, and the linear collider have used substantial CPU time on the farms. Various theory calculations have been accomplished on the farms. Beams simulation calculations have also occasionally been done on these farms. More detailed views of CPU utilization can be found in figures 5-10.

Table 3 shows the total CPU usage of all the farms for the past 13 years by experiment and the sum. D0 leads, followed by CDF. This reflects the current processing requirements of those two experiments. E871 and KTeV are next. Both experiments took substantial amounts of data in the 1999 run. This utilization represents the processing and analysis of that data, as well as Monte Carlo generation. The rest of the users are much smaller, representing the fact that many were active years ago when the amount of CPU was quite small or had or have small amounts of data compared to CDF and D0. This gives a doubling time of 1.019 years. This doubling is driven by the lab's physics program but of course is heavily influenced by the amount of computing that can be afforded. So there is a heavy coupling to Moore's law, and one would expect that the increase is reasonably close to the doubling period of 18 months expected from Moore's law. The reason for a different doubling time is closely related to the fact that the physics programs is not steady – large data runs occur at specific intervals and the computing required is coupled to those periods and is not purchased steadily. Recent purchases have pushed the installed capacity and delivered CPU tremendously – partially skewing this analysis. The initial numbers are also low, given the relatively quiet period at that time.

Farms Configuration

The farms at the present time have the following components of CPU type(s) and number of each type. It is interesting to note that these farms contain twice as many processors as machines, for a total of more than 1000 CPU processors. The largest farms of the past (Run 1 vintage) consisted of approximately 300 single-processor systems.

Table 2. PC Farms Rollup for Fermilab, January 4, 2003

	PC-500	PC-750	PC-800	PC-1000	PC-1260	PC-1670	Spec-INT95	Total PC's
CDF	50		23	64	32	32	18796.8	201
D0	20	50		32		240	45498	342
General	50			40	12		7297.6	90
							71592.4	**633**

Allocations for users are handled by the batch system FBSNG and/or by the physical configuration of the farm. CDF and D0 are dedicated farms. The general farm is allocated dynamically based on quotas and priorities. This system has worked very well and allows projects or experiments to use large fractions of the farm when they have a set of jobs to run. Only when all experiments or projects process simultaneously is there serious contention for the resources.

In addition to CPU there is a large amount of disk space available on the farms. This disk is managed by dfarm and is used as a cache for files as they flow into and out of the farm, as well as a convenient space for storage of intermediate results that are later combined or concatenated in some way to produce files which are then moved off the farms to mass storage or other permanent storage. The current size of dfarm for the farms is given in Table 3. This is not a small amount of disk storage! This is all relatively inexpensive IDE disk and is normally purchased as part of each PC.

	Table 3. Disk Capacity of Farms		
Farm	Total Capacity (TB)	Scratch (TB)	Dfarm (TB)
CDF	16.8	4.1	13.4
D0	16.5	5.4	4.9
General	7.3	2.2	2.0

Plans

The farms are large and will continue to grow as the needs dictate. For these 3 farms the growth is driven by the CPU required for certain special applications. For CDF and D0 the CPU available needs to match the reconstruction and reprocessing and

possibly the Monte Carlo needs of the experiments. In addition, older systems will be decommissioned as they become too old. Already most of the 500 MHz machines have been removed from the CDF and D0 farms and reallocated to other needs. The general farm will grow enough to meet the anticipated demands from the laboratory's program. The 500 MHz machines will be decommissioned.

The big change in farms is the addition of analysis farms to CDF, D0, FNALU and CMS computing systems. These farms are often larger than the production farms and have much more complicated data and user demands. This note does not consider those farms, but they will be the subject of much more scrutiny and attention. The longer-term question of how the analysis farms and reconstruction farms work together (or don't) and how they will be managed is not answered here. Future versions of this note likely will include information about all PC farms at Fermilab. The grid is another component of the farms futures. This is another area where research has started and the status of this will become more clear in the future.

References

1. Enstore, http://hppc.fnal.gov/enstore/
2. dcache, http://www.fnal.gov/docs/products/enstore/
3. dfarm, http://www-isd.fnal.gov/dfarm
4. FBSNG, http://www-isd.fnal.gov/fbsng/
5. NGOP, http://cddocs.fnal.gov/cfdocs/productsDB/ProdDetail.CFM?ProdNum=PU0447

Table 3. Farms Usage: 1991 through 2002.
(SpecInt95-years)

|	1991	1992	1993	1994	1995	1996	1997	1998	1999	2000	2001	2002	Total
D0 offline		4	36	67	74	109	6			194	556	6,723	7,769
CDF		7	20	27	47	60				142	503	1,734	2,540
E871							12	85	209	413	320	518	1,556
KTeV										234	327	505	1,066
BTeV										205	208	60	473
Auger												449	449
NUMI								26	2	1	256	161	445
E831							15	375	29				419
E781							19	98		95	193		405
SDSS											97	144	241
E706	2	5	46	62	50	43	1						209
E791		6	77	78	13	1							176
Auger							55	64	52				172
NuTeV									80	86			166
Beams							7	7	16	95			125
E898											28	64	92
Theory							19		36	4	5	2	66
D0 MC	6	10	25	12	7	1							62
E665	1	7	46	6		1							60
E771		6	13	21	14								54
CMS										49			49
MiniBooNE										32			32
E866						3	26	1					30
E789		10	15										25
E687	6	15	2	2									25
Minos								14					14
Recycler							4	8	1				13
LCD												9	9
E872								1	4	2			7
E835								3			1	1	5
E760		3											3
E731	2												2
Magnet								2					2
	18	72	280	276	205	222	185	664	426	1,550	2,494	10,370	16,761

Figure 5. Farms Usage: CDF and D0.
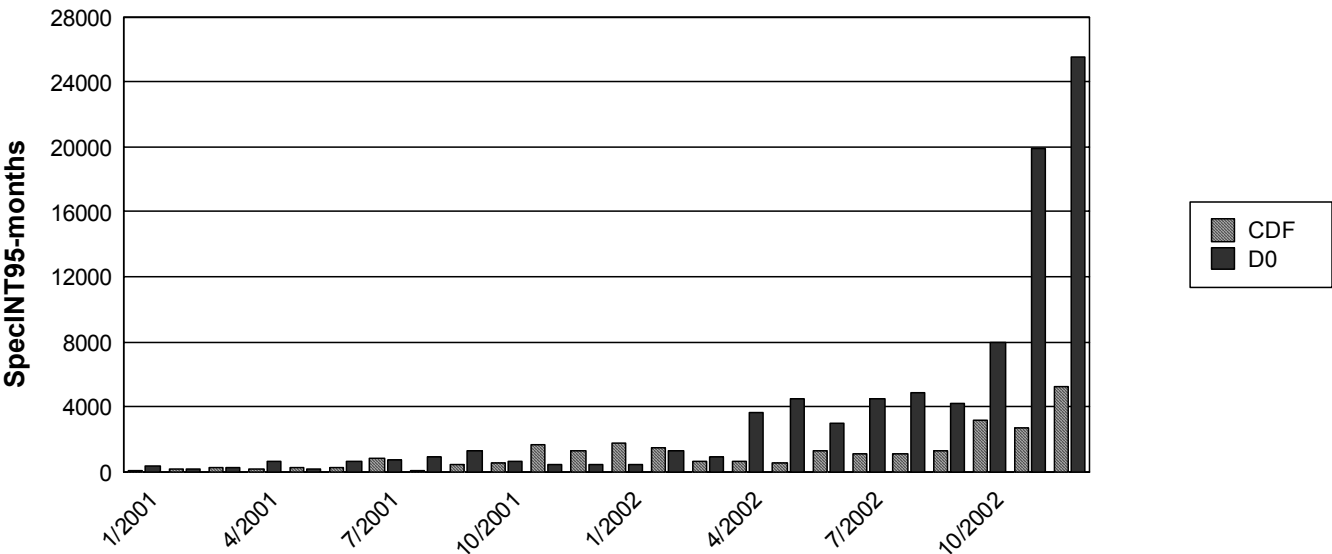
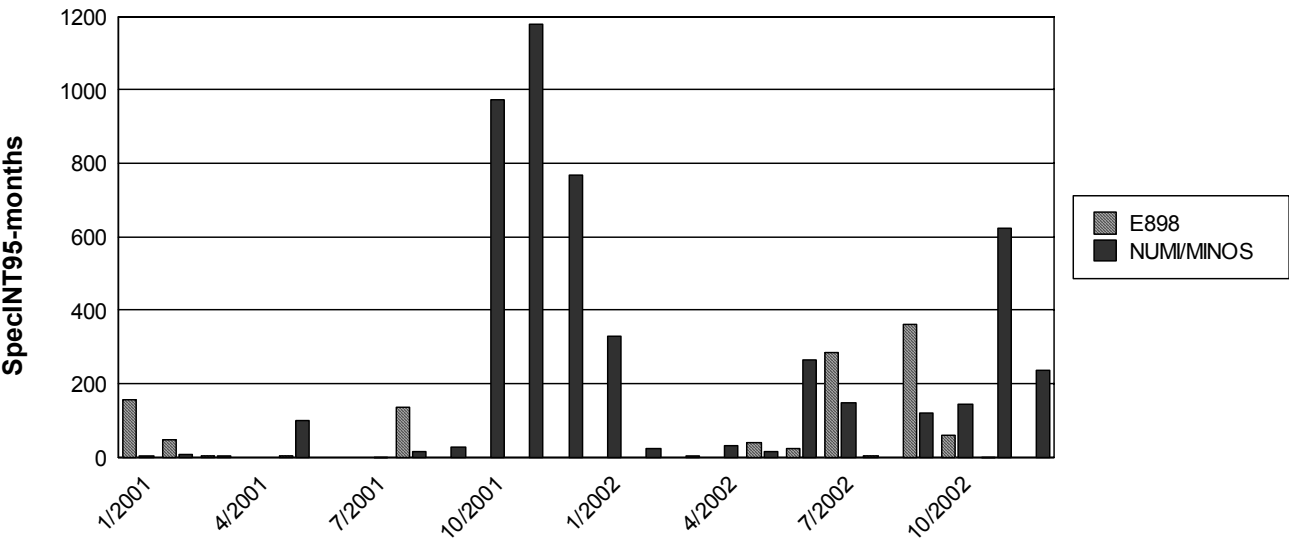Figure 6. Farms Usage: E898 and NUMI/MINOS
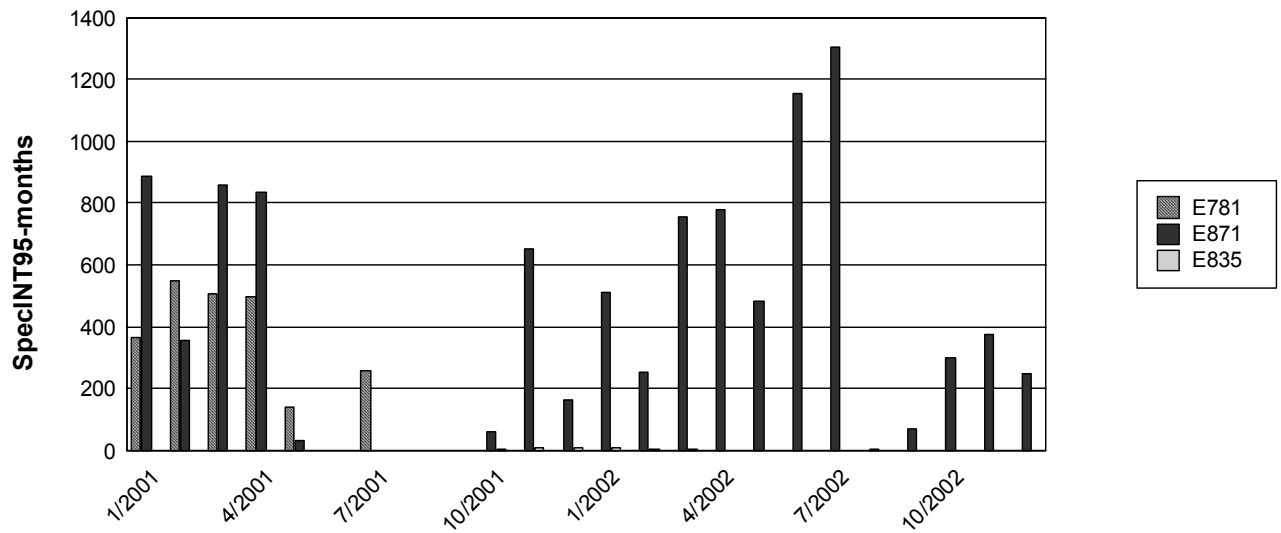
Figure 7. Farms Usage: E781, E835 and E871.
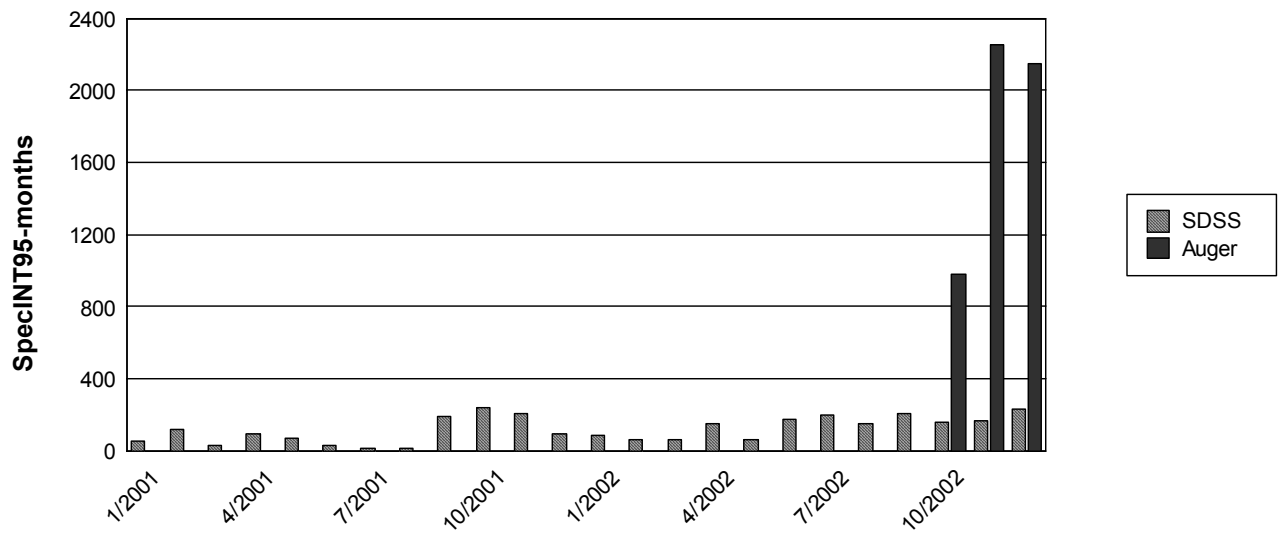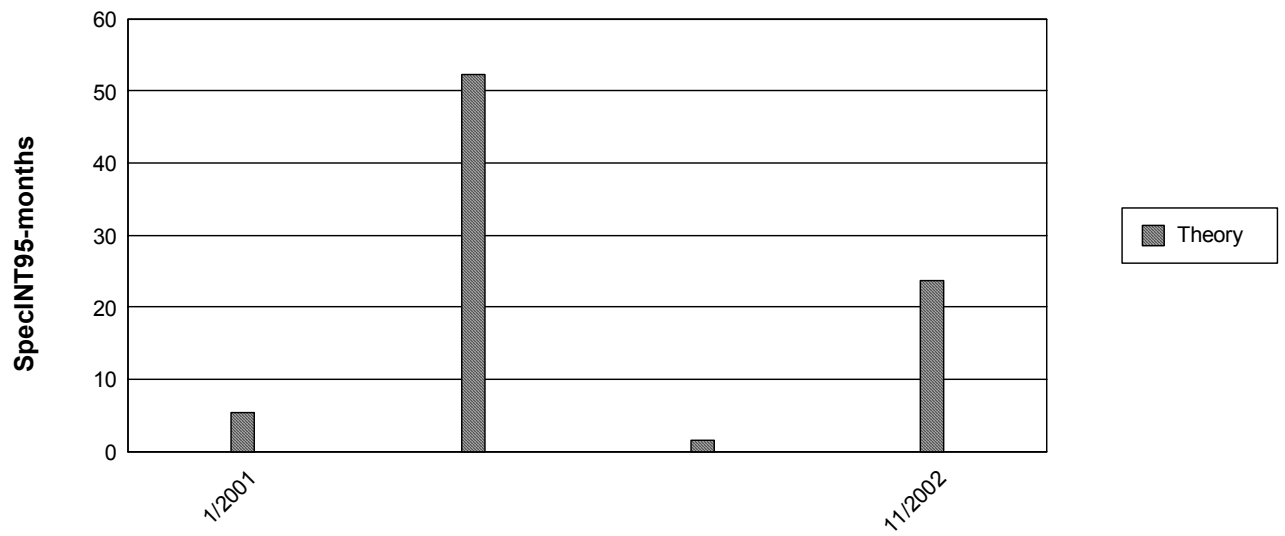

Figure 8. Farms Usage: Auger and SDSS

Figure 9. Farms Usage: Theory.

Figure 10: Farms Usage: BTteV and LCD.